

Tracking Gaze and Visual Focus of Attention of People Involved in Social Interaction

Benoit Massé, Silène Ba, and Radu Horaud

Abstract—The visual focus of attention (VFOA) has been recognized as a prominent conversational cue. We are interested in the VFOA tracking of a group of people involved in social interaction. We note that in this case the participants look either at each other or at an object of interest; therefore they don't always face a camera and, consequently, their gazes (and their VFOAs) cannot be based on eye detection and tracking. We propose a method that exploits the correlation between gaze direction and head orientation. Both VFOA and gaze are modeled as latent variables in a Bayesian switching linear dynamic model. The proposed formulation leads to a tractable learning procedure and to an efficient gaze-and-VFOA tracking algorithm. The method is tested and benchmarked using a publicly available dataset that contains typical multi-party human-robot interaction scenarios, and that was recorded with both a motion capture system, and with a camera mounted onto a robot head.

Index Terms—Visual focus of attention, eye gaze, head pose, dynamic Bayesian model, switching Kalman filter, multi-party dialog, human-robot interaction.

I. INTRODUCTION

Whether engaged in formal meetings or in social gatherings, in addition to speech, people communicate via a large number of non-verbal cues such as prosody, hand gestures, head movements, eye gaze, facial expressions, etc. For example, in a *multi-party* conversation, a common behavior consists of looking either at a person, *e.g.* the speaker, or at an object of current interest, *e.g.* a computer screen, a painting on a wall, or an object lying on a table. This enables participants to respect social etiquette as well as to concentrate their attention onto the current topic of interest. This is also the case in human-computer and human-robot interaction (HRI) scenarios that involve person-to-person, person-to-device and person-to-robot communication.

Among many possible social cues, the *visual focus of attention* (VFOA), or who is looking at whom or at what, has been recognized as one of the most prominent social cues. It is used in multi-party dialog to establish face-to-face communication, to attract someone's attention, or to signify speech-turn taking, thus complementing speech communication. In HRI, a robot (or more generally an intelligent device) must be able to keep temporal knowledge about the locations of the participants, to discriminate between the speaker and the listeners, to recognize the objects of interest, etc.

The VFOA characterizes a perceiver/target pair. It may be defined either by the line from the perceiver's face to the

perceived target, or by the perceiver's *direction of sight* or *gaze direction* (which is often referred to as eye gaze or simply gaze). Indeed, one may state that the VFOA of person i is target j if the perceiver's gaze is aligned with the perceiver-to-target line.

From a physiological point of view, eye gaze depends on both eyeball orientation, and head orientation. Both the eye and the head are rigid bodies with three and six degrees of freedom respectively. The head position (three coordinates) and the head orientation (three angles) are jointly referred to as the *head pose*. With proper choices for the head- and eye-centered coordinate frames, one can assume that gaze is a combination of head pose and of eyeball orientation.¹ Therefore, the VFOA depends on head pose, on eyeball orientation, and on target location.

In this paper we are interested into estimating and tracking jointly the VFOAs of a group of people that communicate with a robot, or *multi-party* HRI, which may well be viewed as a generalization of *single-user* HRI. Nevertheless, from a methodological point of view the former is much more complex than the latter. Indeed, in single-user HRI the person and the robot face each other and hence a camera mounted onto the robot head provides high-resolution frontal images of the user's face such that head pose and eye orientation can both be easily and robustly estimated. In the case of multi-party HRI the eyes can be barely detected since the participants often turn their faces away from the camera. Consequently, VFOA estimation methods based on eye detection and eye tracking are ineffective and one has to estimate the VFOA, indirectly, without explicit eye detection.

We propose a Bayesian switching dynamic model for the estimation and tracking of both gaze directions and of VFOAs of several persons involved in social interaction. While it is assumed that head poses (location and orientation) and target locations can be detected, the unknown gaze directions and VFOAs are treated as latent random variables. The proposed temporal graphical model, that incorporates gaze dynamics and VFOA transitions, leads (i) to a tractable learning algorithm and (ii) to an efficient gaze-and-VFOA tracking method. Moreover, the proposed computational model takes into account a formalism developed in psychophysics [1], [2].

The method is evaluated using the publicly available Vernissage dataset [3]. The dataset consists of recordings of

B. Massé and R. Horaud are with INRIA Grenoble Rhône-Alpes, Montbonnot Saint-Martin, France.

S. Ba is with VideoStitch, Paris, France

This work is supported by ERC Advanced Grant VHIA #340113.

¹Note that orientation generally refers to the pan, tilt and roll angles of a rigid-body pose, while direction refers to the polar and azimuth angles or, equivalently, a unit vector. Since the contribution of the roll angle to gaze is generally marginal, in this paper we make no distinction between orientation and direction.

two persons and a robot that are engaged in situated dialog. During the interaction the participants may gaze to each other, to the robot, or to some wall paintings. The dataset contains two simultaneously collected data, one recorded with a motion capture system (a network of infrared cameras) and one recorded with a camera embedded into the robot head. The motion capture system provides accurate head pose measurements for each participant in the dataset. The ground-truth VFOAs of all the participants were carefully annotated for each frame. The Vernissage dataset therefore allows quantitative evaluation and benchmarking of VFOA estimation in a multi-party HRI scenario.

The remainder of this paper is organized as follows. Section II provides an overview of related work in gaze, VFOA and head-pose estimation. Section III introduces the paper's mathematical notations and definitions, states the problem formulation and describes the proposed model. Section IV presents in detail the model inference and section V derives the learning algorithm. Section VI provides implementation details and Section VII describes the experiments and reports the results.

II. RELATED WORK

As already mentioned, VFOA estimation is closely related to gaze estimation. Hence, many methods have been developed to estimate gaze which is then used to estimate VFOA. In scenarios that rely on precise estimation of gaze direction [4], [5], a head-mounted system, *e.g.* [6], can be used to detect the iris with high accuracy using image processing techniques. Coupled with 3D eye shape extraction, the eyeball orientation can then be easily retrieved. Head-mounted eye trackers provide extremely accurate measurements while they cannot be easily used to estimate the VFOA. Moreover, they are quite pervasive and not appropriate for HRI. Indeed, such devices are likely to disturb the informal nature of the interaction.

Gaze estimation is relevant for a number of scenarios, such as car driving [7] or interaction with smartphones [8]. In these situations, either there is a limited field of view and hence the user performs gaze in a limited range of directions (car driving), or the human is in the loop such that he/she can constantly adapt the position and orientation of the smartphone: frontal images of faces are available in both cases, which in turn provide accurate eye measurements [9], [7], [10], [6]. In some scenarios the user is asked to limit his/her head movements [11], or to proceed through a calibration phase [10], [12]. Even if no specific constraints are imposed, single-user scenarios inherently facilitate the task of eye detection and measurement [9]. At the best of our knowledge (at the time of the writing of this paper), there is no eyeball-based gaze estimation method available, that can deal with unconstrained scenarios, *e.g.* participants that are not facing the cameras, eyes that are partially or totally occluded, etc. Moreover, eye detection and analysis becomes very inaccurate under low-resolution conditions, *e.g.* when participants are faraway from the cameras.

An alternative is to approximate gaze direction with head pose [13]. Unlike eye-based methods, head pose can be estimated from low-resolution images, *i.e.* distant cameras [14],

[15], [16], [17], [18]. Nevertheless, multiple-camera settings are not well suited for HRI. Moreover, these methods estimate gaze only approximatively since eyeball orientation can differ from the head's orientation by $\pm 35^\circ$ [19]. Gaze estimation from head orientation can benefit from the observation that gaze shifts are often achieved by synchronously moving the head and the eyes [20], [1], [2]. The correlation between head pose and gaze has also been exploited in [21].

Several methods were proposed to infer VFOAs either from gaze directions [22], or from head poses [23], [24], [25], [26]. For example, in [26] it is proposed to build a gaze cone around the head orientation and targets lying inside this cone are used to estimate the VFOA. While this method was successfully applied to movies, its limitation resides in its poor accuracy, in particular when two targets are close to each other.

An interesting application of VFOA estimation is the analysis of social behavior of participants engaged in meetings, *e.g.* [21], [27], [23], [28]. Meetings are characterized by interactions between seated people that interact based on speech and on head movements. Some methods estimate the most likely VFOA associated with a head orientation [21], [27]. The drawback of these approaches is that they must be purposely trained for each particular meeting layout. The correlation between VFOA and head pose was also investigated in [23] where an HMM is proposed to infer VFOAs from head and body orientations. This work was extended to deal with more complex scenarios, such as participants interacting with a robot [24], [29]. An input-output HMM is proposed in [29] to incorporate contextual information, *e.g.* participants tend to look to the speaker (another participant or the robot) or to an object which is referred to by a speaker. This drastically improves the performance of VFOA recognition. Nevertheless, it requires additional audio information, such as speaker and speech recognition, or prior information about the scenario itself.

The problem of joint estimation of gaze and of VFOA was addressed in a human-robot cooperation task [25]. In such a scenario the user doesn't necessarily face the camera and robot-mounted cameras have low-resolution, hence the estimation of gaze from direct analysis of eye regions is not sufficiently accurate. [25] proposes to learn a regression between the space of head poses and the space of gaze directions and then to predict an unknown gaze from an observed head pose. The head pose itself is estimated by fitting a 3D elliptical cylinder to a detected face, while the associated gaze direction corresponds to the 3D line joining the head center to the target center. This implies that during the learning stage, the user is instructed to gaze at targets lying on a table in order to provide training data. The regression parameters thus estimated correspond to a discrete set of head-pose/gaze-direction pairs (one for each target), and that an erroneous gaze may be predicted when the latter is not in the range of the set of gaze directions used for training.

This article is an extended version of [30] which summarized the proposed Bayesian dynamic model and reported experiments with motion-capture data. In this paper, we provide a more detailed and comprehensive description of the proposed model, of the learning methodology, and of the associated

algorithm. As mentioned above, in addition to results obtained with motion-capture data, we also provide results using data gathered with an RGB camera mounted onto a robot head. Experimental results and a benchmark conducted with these data show that our method performs well and yields state-of-the-art results.

III. THE PROPOSED MODEL

As already mentioned, the proposed model has its roots in psychophysics [1], [2]. In unconstrained scenarios a person switches his/her gaze from one target to another target, possibly using both eye and head movements. Quick eye movements towards a desired object of interest are called saccades. Eye movements can also be caused by the vestibulo-ocular reflex that compensates for head movements such that one can maintain his/her gaze in the direction of the target of interest. Therefore, in the general case, gazing to an object is achieved by a combination of eye and head movements.

In the case of small gaze shifts, *e.g.* reading or watching TV, eye movements are predominant. In the case of large gaze shifts, often needed in social scenarios, head movements are necessary since eyeball movements have limited range, namely $\pm 35^\circ$. Therefore, the proposed model considers that gaze shifts are produced by head movements that occur simultaneously with eye movements.

A. Problem Formulation

We consider a scenario composed of N active targets and M passive targets. An active target is likely to move and/or to have a leading role in an interaction. Active targets are persons and robots.² Passive targets are objects, *e.g.* a wall painting. The set of all targets is indexed from 0 to $N + M$, where the index 0 corresponds to a non-target. Let i be an active target (a person or a robot), $1 \leq i \leq N$, and j be a passive target (an object), $N + 1 \leq j \leq N + M$. A VFOA is a discrete random variable defined as follows: $V_t^i = j$ means *person (or robot) i looks at target j at time t*. Additionally, the case of a person (or robot) i that looks at none of the targets is defined by $V_t^i = 0$. The case $V_t^i = i$ is excluded. The set of all VFOAs at time t is denoted by $\mathbf{V}_t = (V_t^1, \dots, V_t^N)$.

Two continuous variables are now defined: head orientation and gaze direction. The head orientation of person i at t is denoted with $\mathbf{H}_t^i = [\phi_{H,t}^i, \theta_{H,t}^i]^\top$, *i.e.* the pan and tilt angles of the head with respect to some fixed coordinate frame. The gaze direction of person i is denoted with \mathbf{G}_t^i and is also parameterized by pan and tilt with respect to the same coordinate frame, namely $\mathbf{G}_t^i = [\phi_{G,t}^i, \theta_{G,t}^i]^\top$. Although eyeball orientation is neither needed nor used, it is worth noticing that it is the difference between \mathbf{G}_t^i and \mathbf{H}_t^i .

Finally, to establish a link between VFOAs and gaze directions, the target locations must be defined as well. Let $\mathbf{X}_t^i = [x_t^i, y_t^i, z_t^i]^\top$ be the location of target i . In the case of a person, this location corresponds to the head center while in the case of a passive target, it corresponds to the

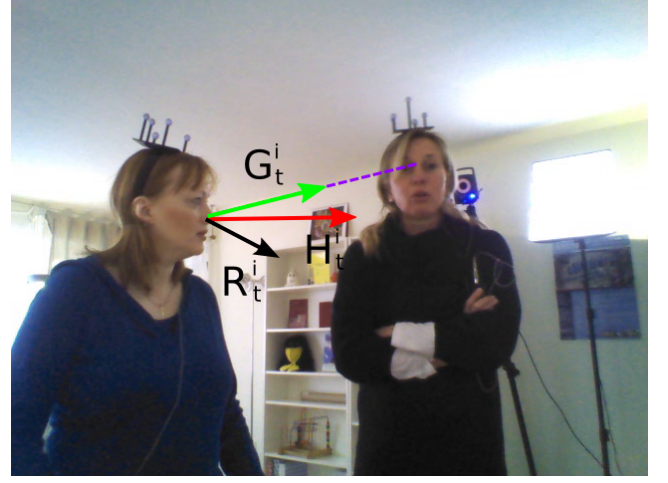


Figure 1. This figure illustrates the main variables associated with a person, *e.g.* the person on the left (indexed i), namely the gaze direction \mathbf{G} (latent), the head orientation \mathbf{H} (observed) and the head reference orientation \mathbf{R} (observed). In this example, the gaze of person i is aligned with \mathbf{X}_t^{ij} (dashed line), which is the direction from person i to the person on the right (indexed j), hence the visual focus of attention of i is equal to j , namely $V_t^i = j$.

target center. These locations are defined in the same coordinate frame as above. Also notice that the direction from the active target i to target j is defined by the unit vector $\mathbf{X}_t^{ij} = (\mathbf{X}_t^j - \mathbf{X}_t^i) / \|\mathbf{X}_t^j - \mathbf{X}_t^i\|$ which can also be parameterized by two angles, $\mathbf{X}_t^{ij} = [\phi_{X,t}^{ij}, \theta_{X,t}^{ij}]^\top$. The variables just defined are illustrated on figure 1.

As already mentioned, target locations and head orientations are observed random variables, while VFOAs and gaze directions are latent random variables. The problem to be solved can now be formulated as a maximum a posteriori (MAP) problem:

$$\hat{\mathbf{V}}_t, \hat{\mathbf{G}}_t = \underset{\mathbf{V}_t, \mathbf{G}_t}{\operatorname{argmax}} P(\mathbf{V}_t, \mathbf{G}_t | \mathbf{H}_{1:t}, \mathbf{X}_{1:t}) \quad (1)$$

Since there is no deterministic relationship between head orientation and gaze direction, we propose to model it probabilistically. For this purpose, we introduce an additional latent random variable, namely the head *reference* orientation, $\mathbf{R}_t^i = [\phi_{R,t}^i, \theta_{R,t}^i]^\top$, which we choose to coincide with the upper-body orientation. We use the following generative model, initially introduced in [23], linking gaze direction, head orientation, and head reference orientation:

$$P(\mathbf{H}_t^i | \mathbf{G}_t^i, \mathbf{R}_t^i; \boldsymbol{\alpha}, \boldsymbol{\Sigma}_H) = \mathcal{N}(\mathbf{H}_t^i; \boldsymbol{\mu}_{H,t}^i, \boldsymbol{\Sigma}_H), \quad (2)$$

$$\text{with } \boldsymbol{\mu}_{H,t}^i = \boldsymbol{\alpha} \mathbf{G}_t^i + (\mathbf{I}_2 - \boldsymbol{\alpha}) \mathbf{R}_t^i, \quad (3)$$

where $\boldsymbol{\Sigma}_H \in \mathbb{R}^{2 \times 2}$ is a covariance matrix, $\mathbf{I}_2 \in \mathbb{R}^{2 \times 2}$ is the identity matrix and $\boldsymbol{\alpha} = \operatorname{Diag}(\alpha_1, \alpha_2)$ is a diagonal matrix of mixing coefficients, $0 < \alpha_1, \alpha_2 < 1$. Also it is assumed that the covariance matrix is the same for all the persons and over time. Therefore, head orientation is an observed random variable normally distributed around a convex combination between two latent variables: gaze direction and head reference orientation.

²Note that in case of a robot, the gaze direction and the head orientation are identical and that the latter can be easily estimated from the head motors.

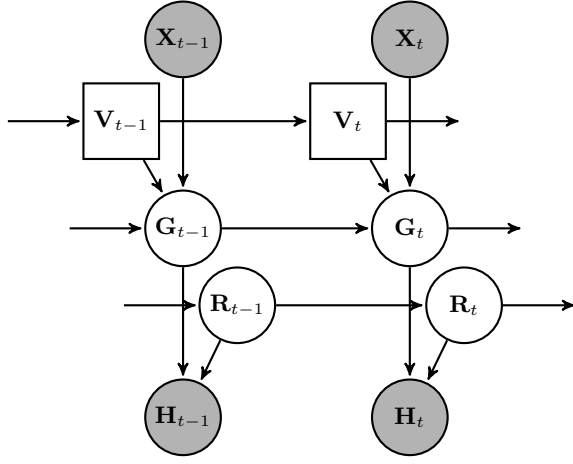


Figure 2. Graphical representation showing the model variables and their dependencies. Squares describe discrete latent variables, circles describe continuous latent variables, and shaded circles describe observations.

B. Gaze Dynamics

The following model is proposed:

$$P(\mathbf{G}_t^i | \mathbf{G}_{t-1}^i, \dot{\mathbf{G}}_{t-1}^i, V_t^i = j, \mathbf{X}_t) = \mathcal{N}(\mathbf{G}_t^i; \boldsymbol{\mu}_{\mathbf{G},t}^{ij}, \boldsymbol{\Gamma}_{\mathbf{G}}), \quad (4)$$

$$P(\dot{\mathbf{G}}_t^i | \dot{\mathbf{G}}_{t-1}^i) = \mathcal{N}(\dot{\mathbf{G}}_t^i; \dot{\mathbf{G}}_{t-1}^i, \boldsymbol{\Gamma}_{\dot{\mathbf{G}}}), \quad (5)$$

with:

$$\boldsymbol{\mu}_{\mathbf{G},t}^{ij} = \begin{cases} \mathbf{G}_{t-1}^i + \dot{\mathbf{G}}_{t-1}^i dt, & \text{if } j = 0, \\ \beta \mathbf{G}_{t-1}^i + (\mathbf{I}_2 - \beta) \mathbf{X}_t^{ij} + \dot{\mathbf{G}}_{t-1}^i dt, & \text{if } j \neq 0, \end{cases} \quad (6)$$

where $\dot{\mathbf{G}}_t^i = d\mathbf{G}_t^i/dt$ is the gaze velocity, $\boldsymbol{\Gamma}_{\mathbf{G}}, \boldsymbol{\Gamma}_{\dot{\mathbf{G}}} \in \mathbb{R}^{2 \times 2}$ are covariance matrices, and $\beta = \text{Diag}(\beta_1, \beta_2)$ is a diagonal matrix of mixing coefficients, $0 < \beta_1, \beta_2 < 1$. Therefore, if a person looks at one of the targets, then his/her gaze dynamics depends on the person-to-target direction \mathbf{X}_t^{ij} at a rate equal to β , and if a person doesn't look at one of the targets, then his/her gaze dynamics follows a random walk.

The head reference orientation dynamics can be defined in a similar way:

$$P(\mathbf{R}_t^i | \mathbf{R}_{t-1}^i, \dot{\mathbf{R}}_{t-1}^i) = \mathcal{N}(\mathbf{R}_t^i; \boldsymbol{\mu}_{\mathbf{R},t}^i, \boldsymbol{\Gamma}_{\mathbf{R}}), \quad (7)$$

$$P(\dot{\mathbf{R}}_t^i | \dot{\mathbf{R}}_{t-1}^i) = \mathcal{N}(\dot{\mathbf{R}}_t^i; \dot{\mathbf{R}}_{t-1}^i, \boldsymbol{\Gamma}_{\dot{\mathbf{R}}}), \quad (8)$$

$$\text{with } \boldsymbol{\mu}_{\mathbf{R},t}^i = \mathbf{R}_{t-1}^i + \dot{\mathbf{R}}_{t-1}^i dt,$$

where $\dot{\mathbf{R}}_t^i = d\mathbf{R}_t^i/dt$ is the head reference orientation velocity and $\boldsymbol{\Gamma}_{\mathbf{R}}, \boldsymbol{\Gamma}_{\dot{\mathbf{R}}} \in \mathbb{R}^{2 \times 2}$ are covariance matrices. The dependencies between all the variables are shown as a graphical model in figure 2.

Given VFOAs along time (see below), we assume that gaze directions, head orientations and head reference orientations are independent. By combining the above equations we obtain:

$$P(\mathbf{H}_t | \mathbf{G}_t, \mathbf{R}_t) = \prod_i \mathcal{N}(\mathbf{H}_t^i; \boldsymbol{\mu}_{\mathbf{H},t}^i, \boldsymbol{\Sigma}_{\mathbf{H}}) \quad (9)$$

$$P(\mathbf{G}_t | \mathbf{G}_{t-1}, \dot{\mathbf{G}}_{t-1}, \mathbf{V}_t, \mathbf{X}_t) = \prod_{i,j} \mathcal{N}(\mathbf{G}_t^i; \boldsymbol{\mu}_{\mathbf{G},t}^{ij}, \boldsymbol{\Gamma}_{\mathbf{G}})^{\delta_j(\mathbf{V}_t^i)} \quad (10)$$

$$P(\mathbf{R}_t | \mathbf{R}_{t-1}, \dot{\mathbf{R}}_{t-1}) = \prod_i \mathcal{N}(\mathbf{R}_t^i; \boldsymbol{\mu}_{\mathbf{R},t}^i, \boldsymbol{\Gamma}_{\mathbf{R}}) \quad (11)$$

where the dependencies between variables are embedded in the variable means, *i.e.* (3) and (6). The covariance matrices will be estimated via training. While gaze directions can vary a lot, head reference orientations are almost constant over time. This can be enforced by

$$\text{Tr}(\boldsymbol{\Gamma}_{\mathbf{G}}) \gg \text{Tr}(\boldsymbol{\Gamma}_{\mathbf{R}}). \quad (12)$$

C. VFOA Dynamics

Using a first-order Markov approximation, the VFOA transition probabilities can be written as:

$$P(\mathbf{V}_t | \mathbf{V}_{1:t-1}) = P(\mathbf{V}_t | \mathbf{V}_{t-1}), \quad (13)$$

Notice that matrix $P(\mathbf{V}_t | \mathbf{V}_{t-1})$ is of size $(N+M)^N \times (N+M)^N$. Indeed, there are N persons (active targets), and $N+M+1$ targets (one "no" target, N active targets and M passive targets) and the case of a person that looks to him/herself is excluded. For example, if $N=2$ and $M=4$, matrix (13) has $(2+4)^{2 \times 2} = 1296$ entries. The estimation of this matrix would require, in principle, a large amount of training data, in particular in the presence of many symmetries. We show that, in practice, only 15 different transitions are possible. This can be seen on the following grounds.

We start by assuming conditional independence between the VFOAs at t :

$$P(\mathbf{V}_t | \mathbf{V}_{t-1}) = \prod_i P(V_t^i | \mathbf{V}_{t-1}). \quad (14)$$

Let's consider V_t^i , the VFOA of person i at t , given \mathbf{V}_{t-1} , the VFOAs at $t-1$. One can distinguish two cases:

- $V_{t-1}^i = k$ where k is either a passive target, $N < k \leq N+M$, or it is none of the targets, $k=0$; in this case V_t^i depends only on V_{t-1}^i , and
- $V_{t-1}^i = k$, where $k \neq i$ is a person $1 \leq k \leq N$; in this case V_t^i depends on the both V_{t-1}^i and V_{t-1}^k .

To summarize, we can write that:

$$P(V_t^i = j | \mathbf{V}_{t-1}) = \begin{cases} P(V_t^i = j | V_{t-1}^i = k, V_{t-1}^k = l) & \text{if } 1 \leq k \leq N, \\ P(V_t^i = j | V_{t-1}^i = k) & \text{otherwise.} \end{cases} \quad (15)$$

Based on this it is now possible to count the total number of possible VFOA transitions. With the same notations as in (15), we have the following possibilities:

- $k=0$ (no target): there are two possible transitions, $j=0$ and $j \neq 0$.
- $N < k \leq N+M$ (passive target): there are three possible transitions, $j=0$, $j=k$, and $j \neq k$.
- $1 \leq k \leq N, l=0$ (active target k looks at no target): there are three possible transitions, $j=0$, $j=k$, and $j \neq k$.
- $1 \leq k \leq N, l=i$ (active target k looks at person i): there are three possible transitions, $j=0$, $j=k$, and $j \neq k$.
- $1 \leq k \leq N, l \neq 0, i$ (active target k looks at active target l different than i): there are four possible transitions, $j=0$, $j=k$, $j=l$ and $j \neq k, l$.

Therefore, there are 15 different possibilities for $P(V_t^i = j | \mathbf{V}_{t-1})$, *i.e.* appendix A. Moreover, by assuming that the

VFOA transitions don't depend on i , we conclude that the transition matrix may have up to 15 different entries. Moreover, the number of possible transitions is even smaller if there is no passive target ($M = 0$), or if the number of active targets is small, *e.g.* $N < 3$. This considerably simplifies the task of estimating this matrix and makes the task of learning tractable.

IV. INFERENCE

We start by simplifying the notation, namely $\mathbf{L}_t = [\mathbf{G}_t; \dot{\mathbf{G}}_t; \mathbf{R}_t; \dot{\mathbf{R}}_t]$ where $[\cdot; \cdot]$ denotes vertical concatenation. The emission probabilities (9) become:

$$P(\mathbf{H}_t | \mathbf{L}_t) = \prod_i \mathcal{N}(\mathbf{H}_t^i; \boldsymbol{\mu}_{\mathbf{H},t}^i, \boldsymbol{\Sigma}_{\mathbf{H}}), \quad (16)$$

$$\text{with } \boldsymbol{\mu}_{\mathbf{H},t}^i = \mathbf{C} \mathbf{L}_t^i, \quad (17)$$

$$\mathbf{C} = (\boldsymbol{\alpha} \quad \mathbf{0} \quad \mathbf{I}_2 - \boldsymbol{\alpha} \quad \mathbf{0}).$$

The 2×8 matrix \mathbf{C} is directly derived from (3). The transition probabilities can be obtained by gathering (10) and (11) with (5) and (8):

$$P(\mathbf{L}_t | \mathbf{V}_t, \mathbf{L}_{t-1}, \mathbf{X}_t) = \prod_i \prod_j \mathcal{N}(\mathbf{L}_t^i; \boldsymbol{\mu}_{\mathbf{L},t}^{ij}, \boldsymbol{\Gamma}_{\mathbf{L}}) \delta_j(\mathbf{V}_t^i), \quad (18)$$

$$\text{with } \boldsymbol{\mu}_{\mathbf{L},t}^{ij} = \mathbf{A}_t^{ij} \mathbf{L}_{t-1}^i + \mathbf{b}_t^{ij} \quad (19)$$

$$\text{and } \boldsymbol{\Gamma}_{\mathbf{L}} = \begin{pmatrix} \boldsymbol{\Gamma}_{\mathbf{G}} & & & \\ & \boldsymbol{\Gamma}_{\dot{\mathbf{G}}} & & \\ & & \boldsymbol{\Gamma}_{\mathbf{R}} & \\ & & & \boldsymbol{\Gamma}_{\dot{\mathbf{R}}} \end{pmatrix}, \quad (20)$$

where \mathbf{A}_t^{ij} is an 8×8 matrix and \mathbf{b}_t^{ij} is an 8×1 vector. The indices i, j and t cannot be dropped since the transitions depend on \mathbf{X}_t^{ij} from (6).

The MAP problem (1) can now be derived in a Bayesian framework for the VFOA variables:

$$P(\mathbf{V}_t | \mathbf{H}_{1:t}, \mathbf{X}_{1:t}) = \int P(\mathbf{V}_t, \mathbf{L}_t | \mathbf{H}_{1:t}, \mathbf{X}_{1:t}) d\mathbf{L}_t. \quad (21)$$

We propose to study the filtering distribution of the joint latent variables, namely $P(\mathbf{V}_t, \mathbf{L}_t | \mathbf{H}_{1:t}, \mathbf{X}_{1:t})$. Indeed, Bayes rule yields:

$$P(\mathbf{V}_t, \mathbf{L}_t | \mathbf{H}_{1:t}, \mathbf{X}_{1:t}) = \frac{1}{c} P(\mathbf{H}_t | \mathbf{L}_t) P(\mathbf{L}_t, \mathbf{V}_t | \mathbf{H}_{1:t-1}, \mathbf{X}_{1:t}). \quad (22)$$

where c is the normalization evidence. Now we can introduce \mathbf{V}_{t-1} and \mathbf{L}_{t-1} using the sum rule:

$$\begin{aligned} P(\mathbf{L}_t, \mathbf{V}_t | \mathbf{H}_{1:t-1}, \mathbf{X}_{1:t}) &= \sum_{\mathbf{V}_{t-1}} \int P(\mathbf{L}_t, \mathbf{V}_t, \mathbf{L}_{t-1}, \mathbf{V}_{t-1} | \mathbf{H}_{1:t-1}, \mathbf{X}_{1:t}) d\mathbf{L}_{t-1} \\ &= \sum_{\mathbf{V}_{t-1}} \int P(\mathbf{L}_t | \mathbf{V}_t, \mathbf{L}_{t-1}, \mathbf{X}_t) P(\mathbf{V}_t | \mathbf{V}_{t-1}) \\ &\quad \times P(\mathbf{L}_{t-1}, \mathbf{V}_{t-1} | \mathbf{H}_{1:t-1}, \mathbf{X}_{1:t-1}) d\mathbf{L}_{t-1}, \end{aligned} \quad (23)$$

where unnecessary dependencies were removed. Combining (22) and (23) we obtain a recursive formulation in $P(\mathbf{V}_t, \mathbf{L}_t | \mathbf{H}_{1:t}, \mathbf{X}_{1:t})$. However, this model is still intractable without further assumptions. The main approximation used

in this work consists of assuming local independence for the posteriors:

$$P(\mathbf{L}_t, \mathbf{V}_t | \mathbf{H}_{1:t}, \mathbf{X}_{1:t}) \simeq \prod_i P(\mathbf{L}_t^i, \mathbf{V}_t^i | \mathbf{H}_{1:t}, \mathbf{X}_{1:t}). \quad (24)$$

A. Switching Kalman Filter Approximation

Several strategies are possible, depending upon the structure of $P(\mathbf{L}_t, \mathbf{V}_t | \mathbf{H}_{1:t}, \mathbf{X}_{1:t})$. Commonly used strategies to evaluate this distribution include variational Bayes or Monte-Carlo. Alternatively, we propose to cast the problem into the framework of switching Kalman filters (SKF) [31]. We assume the filtering distribution to be Gaussian,

$$P(\mathbf{L}_t, \mathbf{V}_t | \mathbf{H}_{1:t}, \mathbf{X}_{1:t}) \propto \mathcal{N}(\mathbf{L}_t; \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t). \quad (25)$$

From (24) and (25) we obtain the following factorization:

$$P(\mathbf{L}_t, \mathbf{V}_t | \mathbf{H}_{1:t}, \mathbf{X}_{1:t}) \propto \prod_i \prod_j \mathcal{N}(\mathbf{L}_t^i; \boldsymbol{\mu}_t^{ij}, \boldsymbol{\Sigma}_t^{ij}) \delta_j(\mathbf{V}_t^i). \quad (26)$$

Thus, (23) can be split into N components, one for each active target i :

$$\begin{aligned} P(\mathbf{L}_t^i, \mathbf{V}_t^i = j | \mathbf{H}_{1:t}, \mathbf{X}_{1:t}) &\propto P(\mathbf{H}_t^i | \mathbf{L}_t^i) \\ &\times \sum_{\mathbf{V}_{t-1}} \int \mathcal{N}(\mathbf{L}_t^i; \mathbf{A}_t^{ij} \mathbf{L}_{t-1}^i + \mathbf{b}_t^{ij}) P(\mathbf{V}_t^i | \mathbf{V}_{t-1}) \\ &\times \prod_k \mathcal{N}(\mathbf{L}_{t-1}^i; \boldsymbol{\mu}_{t-1}^{ik}, \boldsymbol{\Sigma}_{t-1}^{ik}) \delta_k(\mathbf{V}_{t-1}^i) d\mathbf{L}_{t-1}^i, \end{aligned} \quad (27)$$

or, after several algebraic manipulations:

$$P(\mathbf{L}_t^i, \mathbf{V}_t^i = j | \mathbf{H}_{1:t}, \mathbf{X}_{1:t}) \propto \sum_k w_{t-1,t}^{ijk} \mathcal{N}(\mathbf{L}_t^i; \boldsymbol{\mu}_t^{ijk}, \boldsymbol{\Sigma}_t^{ijk}). \quad (28)$$

In this expression, $\boldsymbol{\mu}_t^{ijk}$ and $\boldsymbol{\Sigma}_t^{ijk}$ are obtained by performing constrained Kalman filtering on $\boldsymbol{\mu}_{t-1}^{ik}$, $\boldsymbol{\Sigma}_{t-1}^{ik}$ with transition dynamics defined by \mathbf{A}_t^{ij} and \mathbf{b}_t^{ij} , emission dynamics defined by \mathbf{C} and the observation \mathbf{H}_t^i . The weights $w_{t-1,t}^{ijk}$ are defined as $P(\mathbf{V}_{t-1}^i = k | \mathbf{V}_t^i = j, \mathbf{H}_{1:t}, \mathbf{X}_{1:t})$. The constraint come from the fact that $\|\mathbf{G}_t^i - \mathbf{H}_t^i\| < 35^\circ$ and is achieved by projecting the mean as explained in [32].

This can be rephrased as follows: from the filtering distribution at time $t-1$, there are $N+M$ possible dynamics for \mathbf{L}_t^i . The normal distribution at time $t-1$ then becomes a mixture of $N+M$ normal distributions at time t as shown in (28). However, we expect a single Gaussian such as $P(\mathbf{L}_t^i, \mathbf{V}_t^i = j | \mathbf{H}_{1:t}, \mathbf{X}_{1:t}) \propto \mathcal{N}(\mathbf{L}_t^i; \boldsymbol{\mu}_t^{ij}, \boldsymbol{\Sigma}_t^{ij})$. This can be done by moment matching:

$$\boldsymbol{\mu}_t^{ij} = \sum_k w_{t-1,t}^{ijk} \boldsymbol{\mu}_t^{ijk} \quad (29)$$

$$\boldsymbol{\Sigma}_t^{ij} = \sum_k w_{t-1,t}^{ijk} (\boldsymbol{\Sigma}_t^{ijk} + (\boldsymbol{\mu}_t^{ijk} - \boldsymbol{\mu}_t^{ij})(\boldsymbol{\mu}_t^{ijk} - \boldsymbol{\mu}_t^{ij})^\top) \quad (30)$$

Finally, it is necessary to evaluate $w_{t-1,t}^{ijk}$. Let's introduce the following notations:

$$c_{t-1,t}^{ijk} = P(\mathbf{V}_t^i = j, \mathbf{V}_{t-1}^i = k | \mathbf{H}_{1:t}, \mathbf{X}_{1:t}), \quad (31)$$

$$c_t^{ij} = P(\mathbf{V}_t^i = j | \mathbf{H}_{1:t}, \mathbf{X}_{1:t}). \quad (32)$$

It follows that

$$c_t^{ij} = \sum_k c_{t-1,t}^{ijk} \quad \text{and} \quad w_{t-1,t}^{ijk} = \frac{c_{t-1,t}^{ijk}}{c_t^{ij}}.$$

By applying Bayes formula to $c_{t-1,t}^{ijk}$, yields:

$$c_{t-1,t}^{ijk} \propto P(\mathbf{H}_t^i | V_t^i = j, V_{t-1}^i = k, \mathbf{H}_{1:t-1}, \mathbf{X}_{1:t}) \\ \times c_{t-1}^{ik} P(V_t^i = j | V_{t-1}^i = k, \mathbf{H}_{1:t-1}, \mathbf{X}_{1:t-1}) \quad (33)$$

Then, c_{t-1}^{ik} is obtained from $c_{t-2,t-1}^{ijk}$ calculated at previous time step. The last factor in (33) is either equal to $\sum_l c_{t-1}^{kl} P(V_t^i = j | V_{t-1}^i = k, V_{t-1}^k = l)$ if k is an active target, or $P(V_t^i = j | V_{t-1}^i = k)$ otherwise. Both cases are straightforward to compute. Finally, the first factor in (33), the observation component, can be factorized as $P(\mathbf{H}_t^i | V_t^i = j, V_{t-1}^i = k, \mathbf{H}_{1:t-1}, \mathbf{X}_{1:t}) \times \prod_{n \neq i} \sum_m \sum_p P(\mathbf{H}_t^n | V_t^n = m, V_{t-1}^n = p, \mathbf{H}_{1:t-1}, \mathbf{X}_{1:t})$. By introducing the latent variable \mathbf{L} , we obtain:

$$P(\mathbf{H}_t^n | V_t^n = m, V_{t-1}^n = p, \mathbf{H}_{1:t-1}, \mathbf{X}_{1:t-1}) \\ = \int P(\mathbf{H}_t^n | \mathbf{L}_t^n) P(\mathbf{L}_t^n | \mathbf{L}_{t-1}^n, V_t^n = m) \\ \times P(\mathbf{L}_{t-1}^n | V_{t-1}^n = p, \mathbf{H}_{1:t-1}, \mathbf{X}_{1:t-1}) d\mathbf{L}_{t-1}^n d\mathbf{L}_t^n. \quad (34)$$

All the factors (34) are normal distributions, hence it integrates in closed-form. In summary, we devised a procedure to estimate an online approximation of the joint filtering distribution of the VFOAs, \mathbf{V}_t , and of the gaze and head reference directions, \mathbf{L}_t .

V. LEARNING

The proposed model has two sets of parameters that must be estimated: the transition probabilities associated with the discrete VFOA variables, and the parameters associated with the Gaussian distributions. Learning is carried out using Q annotated recordings, each recording being composed of T_q frames, $1 \leq q \leq Q$. Each recording contains N_q active targets (the robot is the active target 1 and the persons are indexed from 2 to N_q) and M_q passive targets. Both the head poses and the VFOAs of the active targets are available with all the frames.

A. Learning the VFOA Transition Probabilities

The VFOA transitions are drawn from the generalized Bernoulli distribution. Therefore, the transition probabilities can be estimated with $P(V_t^i = j | V_{t-1}^i = k) = \mathbb{E}_{t-1}[\delta_j(V_t^i)]$, where $\delta_j(i)$ is the Kronecker delta function. In section III-C we showed that there are up to 15 different possibilities for the VFOA transition probability. This enables us to derive an explicit formula for each case, see appendix B. Consider for example one of these cases, namely $p_{14} = P(V_t^i = l | V_{t-1}^i = k, V_{t-1}^k = l)$, which is the conditional probability that at t person i looks at target l , given that at $t-1$ person i looked at

person k and that person k looked at target l . This probability can be estimated with the following formula:

$$\hat{p}_{14} = \frac{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{\substack{k=1 \\ k \neq i}}^{N_q} \sum_{\substack{l \neq i, k}} \delta_l(V_t^{q,i}) \delta_k(V_{t-1}^{q,i}) \delta_l(V_{t-1}^{q,k})}{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{\substack{k=1 \\ k \neq i}}^{N_q} \sum_{\substack{l \neq i, k}} \delta_k(V_{t-1}^{q,i}) \delta_l(V_{t-1}^{q,k})}$$

B. Learning the Gaussian Parameters

In section IV we described the derivation of the proposed model that is based on SKF. This model requires the parameters (means and covariances) of the Gaussian distributions defined in (16) and (18). Notice however that the mean (17) of (16) is parameterized by α . Similarly, the mean (19) of (18) is parameterized by β . Consequently, the model parameters are:

$$\theta = (\alpha, \beta, \Gamma_L, \Sigma_H), \quad (35)$$

and we remind that α and β are 2×2 diagonal matrices, Γ_L is a 8×8 covariance and Σ_H is a 2×2 covariance, and that we assumed that these matrices are common to all the active targets. Hence the total number of parameters is equal to $2 + 2 + 36 + 3 = 43$.

In the general case of SKF models, the discrete variables are unobserved both for learning and for inference. Here we propose a learning algorithm that takes advantage of the fact that the discrete variables, *i.e.* VFOAs, are observed during the learning process, namely the VFOAs are annotated. We propose an EM algorithm adapted from [33]. In the case of a standard Kalman filter, an EM iteration alternates between a forward-backward pass to compute the expected latent variables (E-step), and between the maximization of the expected complete-data log-likelihood (M-step).

We start by describing the M-step. The complete-data log-likelihood is:

$$\ln P(\mathbf{L}^1, \mathbf{H}^1, \dots, \mathbf{L}^Q, \mathbf{H}^Q | \theta) \\ = \sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \ln P(\mathbf{L}_t^{q,i} | \mathbf{L}_{t-1}^{q,i}, \beta, \Gamma_L) \\ + \sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=1}^{T_q} \ln P(\mathbf{H}_t^{q,i} | \mathbf{L}_t^{q,i}, \alpha, \Sigma_H). \quad (36)$$

By taking the expectation w.r.t. the posterior distribution $P(\mathbf{L}^1, \dots, \mathbf{L}^Q | \mathbf{H}^1, \dots, \mathbf{H}^Q, \theta)$, we obtain:

$$Q(\theta, \theta^{\text{old}}) = \mathbb{E}_{\mathbf{L}^1, \dots, \mathbf{L}^Q | \theta^{\text{old}}} [\ln P(\mathbf{L}^1, \mathbf{H}^1, \dots, \mathbf{L}^Q, \mathbf{H}^Q | \theta)], \quad (37)$$

which can be maximized w.r.t. to the parameters θ , which yields closed-form expressions for the covariance matrices:

$$\Gamma_L = \frac{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \mathbb{E}[(\mathbf{L}_t^{q,i} - \mu_{\mathbf{L},t}^{q,ij})(\mathbf{L}_t^{q,i} - \mu_{\mathbf{L},t}^{q,ij})^\top]}{\sum_{q=1}^Q (N_q - 1)(T_q - 1)} \quad (38)$$

where $\mu_{\mathbf{L},t}^{q,ij} = \mathbf{A}_t^{q,ij} \mathbf{L}_{t-1}^{q,i} + \mathbf{b}_t^{q,ij}$, *i.e.* (19), and:

$$\Sigma_{\mathbf{H}} = \frac{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=1}^{T_q} \mathbb{E}[(\mathbf{H}_t^{q,i} - \mu_{\mathbf{H},t}^{q,i})(\mathbf{H}_t^{q,i} - \mu_{\mathbf{H},t}^{q,i})^\top]}{\sum_{q=1}^Q (N_q - 1)T_q}, \quad (39)$$

where $\mu_{\mathbf{H},t}^{q,i} = \mathbf{C} \mathbf{L}_t^{q,i}$, *i.e.* (17).

The estimation of α and of β is carried out in the following way. $\partial Q(\theta, \theta^{\text{old}})/\partial \beta_1 = 0$ and $\partial Q(\theta, \theta^{\text{old}})/\partial \beta_2 = 0$ yield a set of two linear equations in the two unknowns:

$$\begin{aligned} \sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \mathbb{E} \left[(\mathbf{L}_t^{q,i} - \mu_{\mathbf{L},t}^{q,ij})^\top \Gamma_{\mathbf{L}}^{-1} \frac{\partial}{\partial \beta_1} (\mathbf{L}_t^{q,i} - \mu_{\mathbf{L},t}^{q,ij}) \right] &= 0, \\ \sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \mathbb{E} \left[(\mathbf{L}_t^{q,i} - \mu_{\mathbf{L},t}^{q,ij})^\top \Gamma_{\mathbf{L}}^{-1} \frac{\partial}{\partial \beta_2} (\mathbf{L}_t^{q,i} - \mu_{\mathbf{L},t}^{q,ij}) \right] &= 0, \end{aligned} \quad (40)$$

and similarly:

$$\begin{aligned} \sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=1}^{T_q} \mathbb{E} \left[(\mathbf{H}_t^{q,i} - \mu_{\mathbf{H},t}^{q,i})^\top \Sigma_{\mathbf{H}}^{-1} \frac{\partial}{\partial \alpha_1} (\mathbf{H}_t^{q,i} - \mu_{\mathbf{H},t}^{q,i}) \right] &= 0, \\ \sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=1}^{T_q} \mathbb{E} \left[(\mathbf{H}_t^{q,i} - \mu_{\mathbf{H},t}^{q,i})^\top \Sigma_{\mathbf{H}}^{-1} \frac{\partial}{\partial \alpha_2} (\mathbf{H}_t^{q,i} - \mu_{\mathbf{H},t}^{q,i}) \right] &= 0, \end{aligned} \quad (41)$$

where as above, the expectation is taken w.r.t. to the posterior distribution. Once the formulas above are expanded and once the means $\mu_{\mathbf{L},t}^{q,ij}$ and $\mu_{\mathbf{H},t}^{q,i}$ are substituted with their expressions, the following terms remain to be estimated: $\mathbb{E}[\mathbf{L}_t^{q,i}]$, $\mathbb{E}[\mathbf{L}_t^{q,i} \mathbf{L}_t^{q,i\top}]$ and $\mathbb{E}[\mathbf{L}_t^{q,i} \mathbf{L}_{t-1}^{q,i\top}]$.

The E-step provides estimates of these expectations via a forward-backward algorithm. For the sake of clarity, we drop the superscripts i (active target index) and q (recording index) up to equation (48) below. Introducing the notation $P(\mathbf{L}_t | \mathbf{H}_1, \dots, \mathbf{H}_t) = \mathcal{N}(\mathbf{L}_t; \mu_t, \mathbf{P}_t)$, the forward-pass equations are:

$$\mu_t = \mathbf{A}_t \mu_{t-1} + \mathbf{b}_t + \mathbf{K}_t (\mathbf{H}_t - \mathbf{C}(\mathbf{A}_t \mu_{t-1} + \mathbf{b}_t)) \quad (42)$$

$$\mathbf{P}_t = (\mathbf{I} - \mathbf{K}_t \mathbf{C}) \mathbf{P}_{t-1}, \quad (43)$$

where:

$$\mathbf{P}_{t,t-1} = \mathbf{A}_t \mathbf{P}_{t-1} \mathbf{A}_t^\top + \Gamma_{\mathbf{L}}, \quad (44)$$

$$\mathbf{K}_t = \mathbf{P}_{t,t-1} \mathbf{C}^\top (\mathbf{C} \mathbf{P}_{t,t-1} \mathbf{C}^\top + \Sigma_{\mathbf{H}})^{-1}. \quad (45)$$

The backward pass estimates $P(\mathbf{L}_t | \mathbf{H}_1, \dots, \mathbf{H}_T) = \mathcal{N}(\mathbf{L}_t; \hat{\mu}_t, \hat{\mathbf{P}}_t)$ and leads to

$$\hat{\mu}_t = \mu_t + \mathbf{J}_t (\hat{\mu}_{t+1} - (\mathbf{A}_{t+1} \mu_t + \mathbf{b}_{t+1})), \quad (46)$$

$$\hat{\mathbf{P}}_t = \mathbf{P}_t + \mathbf{J}_t (\hat{\mathbf{P}}_{t+1} - \mathbf{P}_{t+1,t}) \mathbf{J}_t^\top, \quad (47)$$

where:

$$\mathbf{J}_t = \mathbf{P}_t \mathbf{A}_{t+1}^\top (\mathbf{P}_{t+1,t})^{-1}. \quad (48)$$

The expectations are estimated by performing a forward-backward pass over all the persons and all the recordings of the training data. This yields the following formulas:

$$\mathbb{E}[\mathbf{L}_t^{q,i}] = \hat{\mu}_t^{q,i} \quad (49)$$

$$\mathbb{E}[\mathbf{L}_t^{q,i} \mathbf{L}_t^{q,i\top}] = \hat{\mathbf{P}}_t^{q,i} + \hat{\mu}_t^{q,i} \hat{\mu}_t^{q,i\top} \quad (50)$$

$$\mathbb{E}[\mathbf{L}_t^{q,i} \mathbf{L}_{t-1}^{q,i\top}] = \hat{\mathbf{P}}_t^{q,i} \mathbf{J}_{t-1}^{q,i\top} + \hat{\mu}_t^{q,i} \hat{\mu}_{t-1}^{q,i\top} \quad (51)$$

VI. IMPLEMENTATION DETAILS

A. Dataset

The proposed method was evaluated on the Vernissage dataset [3]. The Vernissage scenario can be briefly described as follows, *e.g.* fig. 3: there are three wall paintings, namely the passive targets denoted with o_1 , o_2 , and o_3 ($M = 3$); two persons, denoted *left* and *right*, interact with the robot, hence $N = 3$. The robot plays the role of an art guide, describing the paintings and asking questions to the two persons in front of him. Each recording is split into two roughly equal parts. The first part is dedicated to painting explanation, with a one-way interaction. The second part consists of a quiz, thus illustrating a dialog between the two participants and the robot, most of the time concerning the paintings.

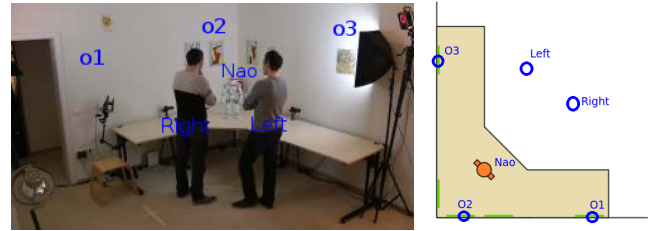


Figure 3. The Vernissage setup. Left: Global view of an “exhibition” showing wall paintings, two participants and the NAO robot. Right: Top view of the room showing the Vernissage layout.

The scene was recorded with a camera embedded into the robot head and with a VICON motion capture system consisting of a network of infrared cameras, placed onto the walls, and of optical markers, placed onto the robot and people heads. Both were recorded at 25 FPS. There is a total of ten recordings, each lasting ten minutes. The VICON system provided accurate estimates of head positions, $\bar{\mathbf{X}}_{1:T}$ and head orientations, $\bar{\mathbf{H}}_{1:T}$. Head positions and head orientations were also estimated using a head-pose algorithm [34] that was applied to the images gathered with the camera embedded into the robot head. Because the whole setup was carefully calibrated the head positions and orientations obtained with the two devices (VICON and robot-head camera) are represented in the same coordinate frame. Finally, the visual focus of attention of the participants was manually annotated in all the frames of all the recordings.

B. Implementation

The inference procedure is summarized in Algorithm 1. This is basically a filtering procedure. The update step consists of applying the recursive relationship, derived in Section IV, to

Algorithm 1 Inference

```

1: procedure GAZEANDVFOA
2:    $\mathbf{X}_1, \mathbf{H}_1 \leftarrow \text{GETOBSERVATIONS}(time = 1)$ 
3:    $c_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1 \leftarrow \text{INITIALIZATION}(\mathbf{H}_1, \mathbf{X}_1)$ 
4:    $V_1^i \leftarrow \text{argmax}_j c_1^{ij}$ 
5:    $\mathbf{G}_1^i \leftarrow \boldsymbol{\mu}_1^{ij} [1..2]$ 
6:   for  $t = 2..T$  do
7:      $\mathbf{X}_t, \mathbf{H}_t \leftarrow \text{GETOBSERVATIONS}(time = t)$ 
8:      $c_t, \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t \leftarrow \text{UPDATE}(\mathbf{H}_t, \mathbf{X}_t, c_{t-1}, \boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})$ 
9:      $V_t^i \leftarrow \text{argmax}_j c_t^{ij}$ 
10:     $\mathbf{G}_t^i \leftarrow \boldsymbol{\mu}_t^{ij} [1..2]$ 
11:   return  $\mathbf{V}_{1:T}, \mathbf{G}_{1:T}$ 

```

$\boldsymbol{\mu}_t^{ij}$, $\boldsymbol{\Sigma}_t^{ij}$ and c_t^{ij} , using $\boldsymbol{\mu}_t^{ijk}$, $\boldsymbol{\Sigma}_t^{ijk}$ and $c_{t-1,t}^{ijk}$ as intermediate variables. The VFOA is chosen using MAP, given observations up to the current frame, and the gaze direction is the mean of the filtered distribution (the first two components of $\boldsymbol{\mu}_t^{ij}$ are indeed the mean for the pan and tilt gaze angles). This algorithm relies on the input observations. They can be obtained either from the motion capture system, with high accuracy, referred to as *Vicon Data* ($\bar{\mathbf{H}}_{1:T}, \bar{\mathbf{X}}_{1:T}$), or with a head pose estimation method, referred to as *Visual Data* ($\hat{\mathbf{H}}_{1:T}, \hat{\mathbf{X}}_{1:T}$). The Vicon Data that we used are publicly available [3].

We now describe the acquisition of the Visual Data which uses the camera mounted onto the robot head. First we use [35] to detect faces and their bounding boxes which are tracked over time using [36]. Then we extract a HOG descriptor from each bounding box and apply the head pose estimation method described in [34]. This yields $\hat{\mathbf{H}}_t$. The 3D head positions, $\hat{\mathbf{X}}_t$, can then be estimated using the line of sight through the face center; the bounding-box size provides a rough estimate of the head depth along this line of sight.

In all our experiments we assumed that the passive targets are static and their positions are provided in advance. The position of the robot itself is also known in advance and one can easily obtain head orientation estimates from the head-motor readings.

Let's now describe the initialization procedure outlined in Algorithm 2. In a probabilistic framework, parameter initialization is generally addressed by defining an initial distribution, *e.g.* $P(\mathbf{L}_1|\mathbf{V}_1)$. Here, we did not explicitly define such a distribution. Initialization is based on the fact that, with repeated similar observation inputs, the algorithm reaches a steady-state. The initialization algorithm uses a repeated update method with initial observation to provide an estimate of gaze and of reference directions. Consequently, the initial filtering distribution $P(\mathbf{L}_1, \mathbf{V}_1|\mathbf{H}_1, \mathbf{X}_1)$ is implicitly defined as the expected stationary state.

VII. EXPERIMENTAL RESULTS

We start by explaining how the learning was carried out, then we show results obtained with the Vicon and visual data. We compare our method with two other methods and we finish with a short discussion.

We applied the same experimental protocol to the Vicon and visual data. The training and testing experiments used

Algorithm 2 Initialization

```

1: procedure INITIALIZATION( $\mathbf{H}_1, \mathbf{X}_1$ )
2:    $\boldsymbol{\mu}_{in} \leftarrow [\mathbf{H}_1; \mathbf{0}; \mathbf{H}_1; \mathbf{0}]$ 
3:    $\boldsymbol{\Sigma}_{in} \leftarrow \mathbf{I}$ 
4:    $c_{in} \leftarrow \frac{1}{N+M}(\text{Uniform})$ 
5:   while Not Convergence do
6:      $c_{in}, \boldsymbol{\mu}_{in}, \boldsymbol{\Sigma}_{in} \leftarrow \text{UPDATE}(\mathbf{H}_1, \mathbf{X}_1, c_{in}, \boldsymbol{\mu}_{in}, \boldsymbol{\Sigma}_{in})$ 
7:   return  $c_{in}, \boldsymbol{\mu}_{in}, \boldsymbol{\Sigma}_{in}$ 

```

the *leave-one-out* strategy. We used the frame recognition rate (FRR) as a metric to quantitatively evaluate the method. FRR evaluates the percentage of frames for which the VFOA is correctly estimated. One should note however that the ground-truth VFOAs were obtained by manually annotating each frame in the data. This is subject to errors because the annotator has to associate a target with a person. Obviously this is prone to errors.

The VFOA transition probabilities and the model parameters were estimated using the learning method described in section V. Appendix B provides the formulas used for estimating the VFOA transition probabilities given the annotated data. Notice that the 15 transitions probabilities thus estimated are shared by the Vicon and the visual data.

The Gaussian parameters, *i.e.* (35), were estimated using the EM algorithm of section V-B. This learning procedure requires head-pose estimates as well as the transition probabilities, estimated as just explained. Since these estimates are different for the two kinds of data (Vicon and visual) we carried out the learning process twice, using the Vicon data and the visual data. The EM algorithm needs initialization. The initial parameter values for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are

$$\boldsymbol{\alpha}^0 = \boldsymbol{\beta}^0 = \text{Diag}(0.5, 0.5).$$

The covariance matrices $\boldsymbol{\Sigma}_H$ and $\boldsymbol{\Gamma}_L$ defined in (20) are initialized with

$$\begin{aligned} \boldsymbol{\Sigma}_H^0 &= \sigma \mathbf{I}_2, & \text{with } \sigma &= 15 \\ \boldsymbol{\Gamma}_G^0 &= \boldsymbol{\Gamma}_G^0 = \gamma \mathbf{I}_2, & \text{with } \gamma &= 5 \\ \boldsymbol{\Gamma}_R^0 &= \boldsymbol{\Gamma}_R^0 = \eta \mathbf{I}_2, & \text{with } \eta &= 0.5 \end{aligned}$$

A. Vicon Data

The FRR of the estimated VFOAs for the Vicon data are summarized in table I. A few examples are shown in figure 5. The FRR score varies between 28.3% and 74.4% for [23] and between 43.1% and 79.8% for the proposed method. Notice that high scores are obtained by both methods for recording #27. Similarly, low scores are obtained for recording #26. Since both methods assume that head motions and gaze shifts occur synchronously, an explanation could be that this hypothesis is well suited for some participants.

The confusion matrices for VFOA classification using Vicon data are given in figure 4. A few similarities between both methods appear. In particular, wall painting #0₂ stands just behind Nao and both methods regularly confuse these two targets. In addition, the head of one of the persons is often aligned with painting #0₁ from the viewpoint of the

Table I
FRR SCORES OF THE ESTIMATED VFOAs FOR THE VICON DATA.

Recording	Ba [23]		Proposed	
	left	right	left	right
09	51.6	65.1	59.8	61.4
10	64.3	74.4	76.5	65.0
12	53.5	67.6	61.6	63.2
15	67.1	46.2	64.8	67.6
18	37.5	28.3	62.0	53.7
19	56.7	45.4	54.5	60.4
24	44.9	49.0	59.7	54.7
26	40.3	32.9	43.6	43.1
27	65.8	72.0	79.8	78.3
30	69.1	49.1	72.0	63.9
Mean	54.5		62.6	

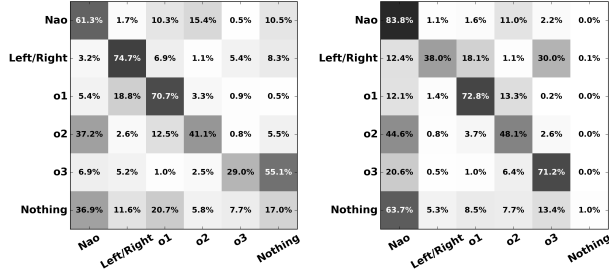


Figure 4. Confusion matrices for [23] (left) and for the proposed method (right) using the Vicon data. Rows: ground-truth VFOA. Columns: estimated VFOA. Diagonal entries represent the recall.

other person. A similar remark holds for painting # o_3 . As a consequence both methods often confuse the VFOA in these cases. This can be seen in the third image of figure 5. Indeed, it is difficult to estimate whether person *left* looks either at # o_1 or person *right*.

Finally, both methods have problems with recognizing the VFOA “nothing” or gaze aversion ($V_t^i = 0$). We propose the following explanation: the targets are widespread in the scene, hence it is likely that an acceptable target lies in most of the gaze directions. Moreover, Nao is centrally positioned, therefore the head orientation used to look at Nao is similar to the resting head orientation used for gaze aversion. However, in [23] the reference head orientation is fixed and poorly suited for dynamic head-to-gaze mapping, hence the high error rate on painting # o_3 . Our method favors the selection of a target, either active or passive, over the no-target case.

B. Visual Data

The head pose estimation method [34] was applied to the visual data, thus yielding observations (head orientation and head position) for our method. Table II shows the accuracy of these measurements (in degrees and in centimeters), when compared with the motion capture data.

As it can be seen in this table, while the head orientation is quite accurate, the head position varies a lot. In particular, for recordings #19 and #24, the head position errors are roughly 80 cm, while the participants are in between 1.5 m and 2.5 m in front of the robot. This is because the head positions are estimated from the size of the bounding boxes, and the size of

Table II
MEAN ERROR FOR HEAD POSE ESTIMATIONS FOR THE *left* AND *right* PERSONS. THE ESTIMATED HEAD POSITIONS (CENTIMETERS) AND ORIENTATIONS (DEGREES) ARE COMPARED WITH THE MOTION CAPTURE DATA AS PROVIDED BY THE VICON SYSTEM.

Recording	Head position (cm)		Head Pan		Head Tilt	
	left	right	left	right	left	right
09	18.1	20.8	4.4°	4.8°	3.7°	3.2°
12	35.7	41.5	4.8°	5.5°	2.6°	3.8°
18	36.9	12.8	6.8°	3.7°	5.8°	2.5°
19	86.0	87.4	4.0°	5.8°	2.7°	3.7°
24	86.5	73.9	3.3°	3.5°	2.8°	2.7°
26	50.2	56.9	7.4°	9.0°	4.1°	5.2°
27	64.5	58.3	4.1°	5.8°	3.2°	4.4°
30	16.7	13.3	2.8°	2.9°	1.8°	2.7°
Mean	46.4		5.0°		3.3°	

these boxes varies. In particular the error in head position is larger as the participants are farther away from the robot. In these cases, the bounding box is larger than it should be and hence the head position is, on an average, one meter closer than the true position. The error in head pose estimation affects the quality of the results.

Table III
FRR SCORES OF THE ESTIMATED VFOAs FOR THE VISUAL DATA.

Recording	Ba [23]		Proposed	
	left	right	left	right
09	50.3	59.8	58.1	55.9
12	54.2	14.8	59.0	46.5
18	39.0	16.1	64.2	33.1
27	38.2	17.1	53.3	55.1
30	61.6	44.6	54.7	66.6
Mean	39.0		54.7	

The FRR scores using the Visual data are shown in table III. As expected the loss in accuracy is highly correlated with the error in head position: the results obtained with recordings #09 and #30 are close to the ones obtained with the Vicon data, whereas there is a significant loss in accuracy for the other recordings. The loss is notable for [23] in the case of the *right* person in recordings #12, #18 and #27.

The confusion matrices obtained with the Visual data are shown in figure 6. The patterns observed in the case of the Vicon data are also present in visual data experiments.

Unfortunately, we were unable to thoroughly compare our method with [29] for several reasons. The results reported in [29] with the visual data were obtained with an unpublished in-house head-pose detection and tracking method. Since both head position and head orientation inputs play a fundamental role in gaze estimation, the comparison between our method and [29] is somehow biased. Moreover, [29] uses contextual information, namely the identity of the speaker (one of the participants or the robot) as well as the object of interest. Finally, we note that in [29] only mean FRR values obtained over all the recordings are reported.

Table IV summarizes a comparison between the average FRR obtained with our method and with [23] and [29]. We note that, overall, our method yields a similar FRR score as [29] using the Vicon data (first row). In this case the two methods share the same head pose inputs provided by

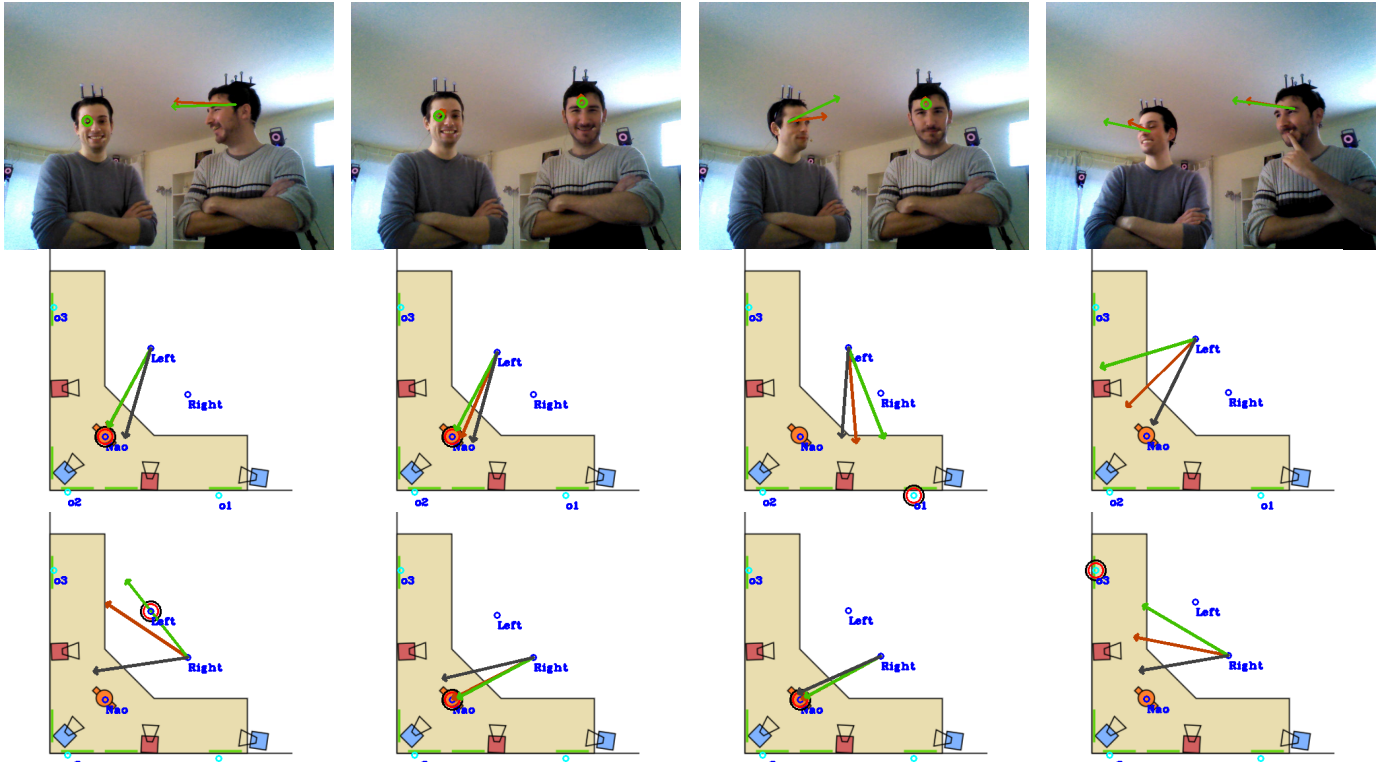


Figure 5. Results obtained with the proposed method on Vicon data. Gaze directions are shown with green arrows, head reference directions with dark-grey arrows and observed head directions with red arrows. The ground-truth VFOA is shown with a black circle. The top row displays the image of the robot-head camera. Top views of the room show results obtained for the Left (middle row) and Right (bottom row) persons. In the last example the Left person gazes at “nothing”.

Nao	44.6%	2.3%	14.1%	20.5%	1.3%	16.8%
Left/Right	3.9%	39.5%	21.4%	4.3%	20.1%	10.5%
o1	11.9%	14.6%	62.0%	8.6%	0.2%	2.5%
o2	39.3%	3.0%	18.9%	27.3%	0.8%	10.5%
o3	9.8%	2.8%	5.2%	5.1%	31.6%	45.4%
Nothing	46.8%	8.0%	13.6%	8.8%	11.3%	11.3%
	Nao	Left/Right	o1	o2	o3	Nothing

Nao	74.9%	6.1%	1.8%	14.3%	2.7%	0.0%
Left/Right	26.1%	32.2%	10.6%	4.3%	26.7%	0.1%
o1	16.8%	8.1%	48.2%	25.4%	1.3%	0.0%
o2	48.8%	3.4%	3.0%	41.8%	2.7%	0.0%
o3	29.6%	1.1%	4.8%	9.4%	54.8%	0.0%
Nothing	68.1%	8.3%	2.5%	7.7%	13.2%	0.0%
	Nao	Left/Right	o1	o2	o3	Nothing

Figure 6. Confusion matrix for [23] (left) and our method (right) on visual data. Rows: ground-truth VFOA. Columns: estimated VFOA. Diagonal terms represent the recall.

the motion capture system. [29] uses additional contextual information that is not used by our method. In the case of visual data (second row) the difference in performance between the two methods may be explained by the fact that [29] uses a head-pose tracker that yields a smooth head trajectory and that filters our noisy detections.

Table IV
MEAN FRR SCORES OBTAINED WITH [23], WITH [29] AND WITH THE PROPOSED METHOD. RECORDING #26 WAS EXCLUDED FROM THE FRR MEANS AS REPORTED IN [29]. MOREOVER, [29] USES ADDITIONAL CONTEXTUAL INFORMATION.

	Ba [23]	Sheikhi [29]	Proposed
Vicon data	56.5	66.6	64.7
Visual data	39.0	62.4	54.7

VIII. CONCLUSIONS

In this paper we addressed the problem of detecting and tracking the visual focus of attention of a group of persons involved in social interaction. We proposed a probabilistic formulation that exploits the correlation between head orientation and gaze on one side, and between visual focus of attention and gaze on the other side. We described in detail the proposed model. In particular we showed that the entries of the large-sized matrix of VFOA transition probabilities have a very small number of different possibilities for which we provided closed-form formulae. The immediate consequence of this simplified transition matrix is that the associated learning doesn't require a large training dataset. We showed that the problem of simultaneously inferring VFOAs and gaze directions over time can be cast in the framework of a switching linear dynamic model, which yields a tractable inference algorithm.

We applied the proposed method to the Vernissage dataset that contains recordings of a human-robot interaction scenario. We experimented both with motion-capture data gathered with a Vicon system and with visual data gathered with a camera mounted onto a robot head. We compared our method with two other methods, one based on HMM [23], and one based on input-output HMM [29]. While both [23] and our method use exactly the same input, [29] uses a different head-pose estimator along with contextual information.

We acknowledge that contextual information about the scenario at hand can considerably improve the results. Within

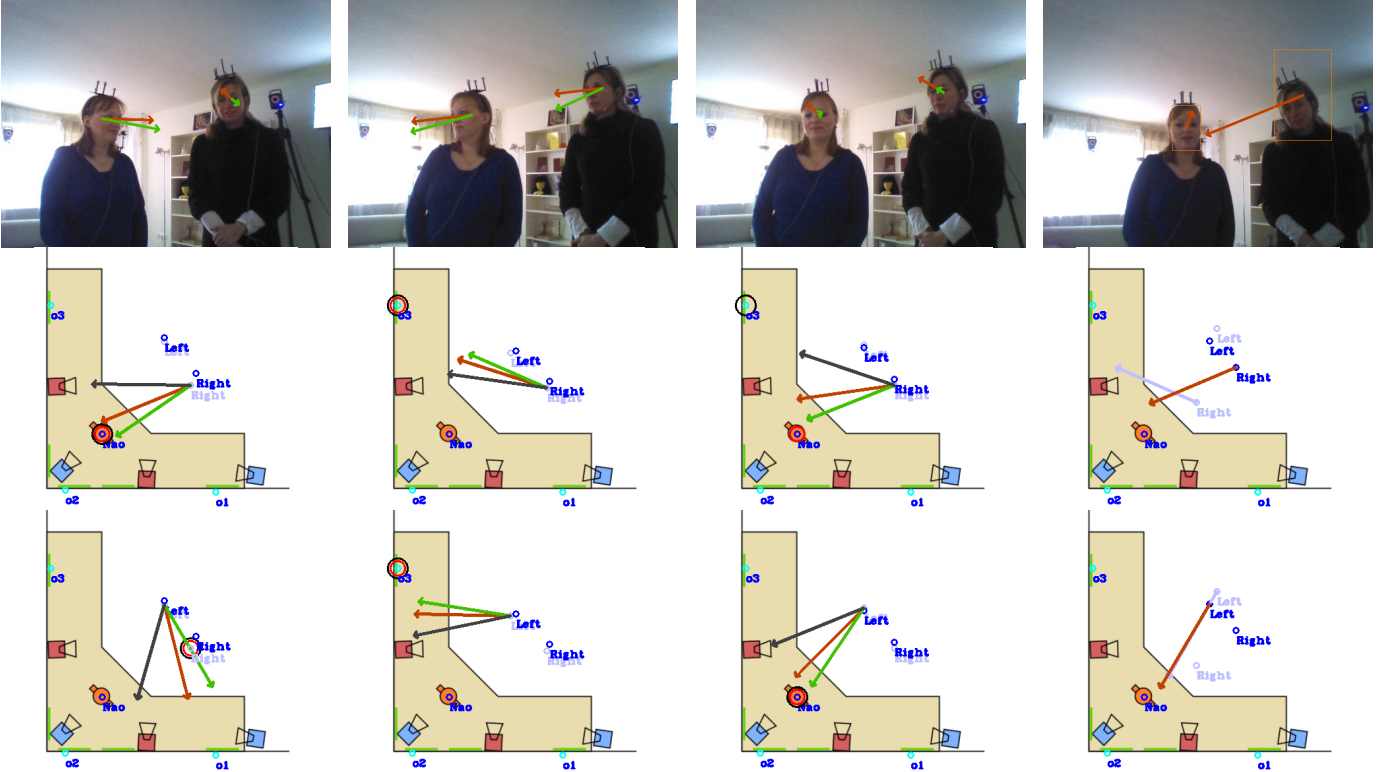


Figure 7. Results obtained with the proposed method on visual data. Gaze directions are shown with green arrows, head reference directions with dark-grey arrows and observed head directions with red arrows. The ground-truth VFOA is shown with a black circle. The top row displays the image of the robot-head camera. Top views of the room show results obtained for the Left (middle row) and Right (bottom row) persons. The last column shows an error in tracking that misleads the head pose estimator for the Right person.

the framework of the proposed method such additional information, *e.g.* speech-turn information, could be easily plugged into the dynamics of the interaction. For example, speaker recognition and localization and speech recognition may be used to learn VFOA transitions for multi-party multimodal dialog systems.

We note that gaze inference from head orientation is an ill-posed problem. Indeed, the correlation between gaze and head movements depends from one person to another. It is however a necessary process whenever the eyes cannot be reliably extracted from images and analyzed. We proposed to guide this inference based on detecting alignments between gaze directions and a finite number of targets often associated with social interaction.

As with any machine learning formalism, the effectiveness of the method depends on the quality of an annotated training dataset. In the case of the Vernissage dataset, the VFOAs (target j that is gazed by person i , for each person and in every frame) were manually annotated. Clearly these annotations are prone to errors because there is no quantitative way to measure the actual VFOA of a person whose eyes are not visible. This explains some of the large errors between the manually annotated VFOAs and the estimated VFOAs, both with the Vicon and the visual data.

In the future we plan to investigate discriminative methods based on neural network architectures for inferring gaze directions from head orientations in the presence of a finite number of targets. For example one could train a deep learning

network from pairs of head orientations and gaze directions. This implies that a large training dataset is available, and we plan to collect, annotate and release such a dataset. One can, for instance use a multiple-camera setup to be able to detect the eyes of several participants and to estimate their head poses, as well as a microphone array in order to robustly extract speech information.

APPENDIX A VFOA TRANSITION PROBABILITIES

Using the notations introduced in section III-C, let i , $1 \leq i \leq N$, be an active target. In section III-C we showed that in practice the probability transition matrix has up to 15 different entries. For completeness, these entries are listed below.

- The VFOA of i at $t - 1$ is neither an active nor a passive target ($k = 0$):

$$p_1 = P(V_t^i = 0 | V_{t-1}^i = 0)$$

$$p_2 = P(V_t^i = j | V_{t-1}^i = 0)$$

- The VFOA of i at $t - 1$ is a passive target ($N < k \leq N + M$):

$$p_3 = P(V_t^i = 0 | V_{t-1}^i = k)$$

$$p_4 = P(V_t^i = k | V_{t-1}^i = k)$$

$$p_5 = P(V_t^i = j | V_{t-1}^i = k)$$

- The VFOA of i at $t-1$ is an active target ($1 \leq k \leq N, k \neq i$):

$$p_6 = P(V_t^i = 0 | V_{t-1}^i = k, V_{t-1}^k = 0)$$

$$p_7 = P(V_t^i = k | V_{t-1}^i = k, V_{t-1}^k = 0)$$

$$p_8 = P(V_t^i = j | V_{t-1}^i = k, V_{t-1}^k = 0)$$

$$p_9 = P(V_t^i = 0 | V_{t-1}^i = k, V_{t-1}^k = i)$$

$$p_{10} = P(V_t^i = k | V_{t-1}^i = k, V_{t-1}^k = i)$$

$$p_{11} = P(V_t^i = j | V_{t-1}^i = k, V_{t-1}^k = i)$$

$$p_{12} = P(V_t^i = 0 | V_{t-1}^i = k, V_{t-1}^k = l)$$

$$p_{13} = P(V_t^i = k | V_{t-1}^i = k, V_{t-1}^k = l)$$

$$p_{14} = P(V_t^i = l | V_{t-1}^i = k, V_{t-1}^k = l)$$

$$p_{15} = P(V_t^i = j | V_{t-1}^i = k, V_{t-1}^k = l)$$

APPENDIX B VFOA LEARNING

This appendix provides the formulae allowing to estimate the 15 transitions probabilities as explained in section V-A.

$$\hat{p}_1 = \frac{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \delta_0(V_t^{q,i}) \delta_0(V_{t-1}^{q,i})}{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \delta_0(V_{t-1}^{q,i})}$$

$$\hat{p}_2 = \frac{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{j \neq i} \delta_j(V_t^{q,i}) \delta_0(V_{t-1}^{q,i})}{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \delta_0(V_{t-1}^{q,i})}$$

$$\hat{p}_3 = \frac{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{k=N_q+1}^{N_q+M_q} \delta_0(V_t^{q,i}) \delta_k(V_{t-1}^{q,i})}{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{k=N_q+1}^{N_q+M_q} \delta_k(V_{t-1}^{q,i})}$$

$$\hat{p}_4 = \frac{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{k=N_q+1}^{N_q+M_q} \delta_k(V_t^{q,i}) \delta_k(V_{t-1}^{q,i})}{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{k=N_q+1}^{N_q+M_q} \delta_k(V_{t-1}^{q,i})}$$

$$\hat{p}_5 = \frac{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{k=N_q+1}^{N_q+M_q} \sum_{j \neq i, k} \delta_j(V_t^{q,i}) \delta_k(V_{t-1}^{q,i})}{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{k=N_q+1}^{N_q+M_q} \delta_k(V_{t-1}^{q,i})}$$

$$\hat{p}_6 = \frac{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{\substack{k=1 \\ k \neq i}}^{N_q} \delta_0(V_t^{q,i}) \delta_k(V_{t-1}^{q,i}) \delta_0(V_{t-1}^{q,k})}{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{\substack{k=1 \\ k \neq i}}^{N_q} \delta_k(V_{t-1}^{q,i}) \delta_0(V_{t-1}^{q,k})}$$

$$\hat{p}_7 = \frac{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{\substack{k=1 \\ k \neq i}}^{N_q} \delta_k(V_t^{q,i}) \delta_k(V_{t-1}^{q,i}) \delta_0(V_{t-1}^{q,k})}{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{\substack{k=1 \\ k \neq i}}^{N_q} \delta_k(V_{t-1}^{q,i}) \delta_0(V_{t-1}^{q,k})}$$

$$\hat{p}_8 = \frac{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{\substack{k=1 \\ k \neq i}}^{N_q} \sum_{j \neq i, k} \delta_j(V_t^{q,i}) \delta_k(V_{t-1}^{q,i}) \delta_0(V_{t-1}^{q,k})}{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{\substack{k=1 \\ k \neq i}}^{N_q} \delta_k(V_{t-1}^{q,i}) \delta_0(V_{t-1}^{q,k})}$$

$$\hat{p}_9 = \frac{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{\substack{k=1 \\ k \neq i}}^{N_q} \delta_0(V_t^{q,i}) \delta_k(V_{t-1}^{q,i}) \delta_i(V_{t-1}^{q,k})}{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{\substack{k=1 \\ k \neq i}}^{N_q} \delta_k(V_{t-1}^{q,i}) \delta_i(V_{t-1}^{q,k})}$$

$$\hat{p}_{10} = \frac{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{\substack{k=1 \\ k \neq i}}^{N_q} \delta_k(V_t^{q,i}) \delta_k(V_{t-1}^{q,i}) \delta_i(V_{t-1}^{q,k})}{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{\substack{k=1 \\ k \neq i}}^{N_q} \delta_k(V_{t-1}^{q,i}) \delta_i(V_{t-1}^{q,k})}$$

$$\hat{p}_{11} = \frac{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{\substack{k=1 \\ k \neq i}}^{N_q} \sum_{j \neq i, k} \delta_j(V_t^{q,i}) \delta_k(V_{t-1}^{q,i}) \delta_i(V_{t-1}^{q,k})}{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{\substack{k=1 \\ k \neq i}}^{N_q} \delta_k(V_{t-1}^{q,i}) \delta_i(V_{t-1}^{q,k})}$$

$$\hat{p}_{12} = \frac{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{\substack{k=1 \\ k \neq i}}^{N_q} \sum_{l \neq i, k} \delta_0(V_t^{q,i}) \delta_k(V_{t-1}^{q,i}) \delta_l(V_{t-1}^{q,k})}{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{\substack{k=1 \\ k \neq i}}^{N_q} \sum_{l \neq i, k} \delta_k(V_{t-1}^{q,i}) \delta_l(V_{t-1}^{q,k})}$$

$$\hat{p}_{13} = \frac{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{k=1}^{N_q} \sum_{l \neq i, k} \delta_k(V_t^{q,i}) \delta_k(V_{t-1}^{q,i}) \delta_l(V_{t-1}^{q,k})}{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{k=1}^{N_q} \sum_{l \neq i, k} \delta_k(V_{t-1}^{q,i}) \delta_l(V_{t-1}^{q,k})}$$

$$\hat{p}_{14} = \frac{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{k=1}^{N_q} \sum_{l \neq i, k} \delta_l(V_t^{q,i}) \delta_k(V_{t-1}^{q,i}) \delta_l(V_{t-1}^{q,k})}{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{k=1}^{N_q} \sum_{l \neq i, k} \delta_k(V_{t-1}^{q,i}) \delta_l(V_{t-1}^{q,k})}$$

$$\hat{p}_{15} = \frac{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{k=1}^{N_q} \sum_{l \neq i, k} \sum_{j \neq i, k, l} \delta_j(V_t^{q,i}) \delta_k(V_{t-1}^{q,i}) \delta_l(V_{t-1}^{q,k})}{\sum_{q=1}^Q \sum_{i=2}^{N_q} \sum_{t=2}^{T_q} \sum_{k=1}^{N_q} \sum_{l \neq i, k} \delta_k(V_{t-1}^{q,i}) \delta_l(V_{t-1}^{q,k})}$$

ACKNOWLEDGMENTS

The authors would like to thank Vincent Drouard for his valuable expertise in head pose estimation and tracking.

REFERENCES

- [1] E. G. Freedman and D. L. Sparks, "Eye-head coordination during head-unrestrained gaze shifts in rhesus monkeys," *Journal of Neurophysiology*, 1997.
- [2] E. G. Freedman, "Coordination of the eyes and head during visual orienting," *Experimental Brain Research*, vol. 190, 2008.
- [3] D. B. Jayagopi *et al.*, "The vernissage corpus: A multimodal human-robot-interaction dataset," IDIAP, Tech. Rep., 2012.
- [4] L. H. Yu and M. Eizenman, "A new methodology for determining point-of-gaze in head-mounted eye tracking systems," *IEEE Transactions on Biomedical Engineering*, vol. 51, Oct 2004.
- [5] T. Toyama, T. Kieninger, F. Shafait, and A. Dengel, "Gaze guided object recognition using a head-mounted eye tracker," in *Proceedings of the ETRA Symposium*, 2012.
- [6] A. K. A. Hong, J. Pelz, and J. Cockburn, "Lightweight, low-cost, side-mounted mobile eye tracking system," in *IEEE WNYIPW*, 2012.
- [7] P. Smith, M. Shah, and N. Da Vitoria Lobo, "Determining driver visual attention with one camera," *IEEE Transactions on Intelligent Transportation Systems*, vol. 4, 2003.
- [8] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusk, and A. Torralba, "Eye tracking for everyone," in *IEEE CVPR*, June 2016.
- [9] Y. Matsumoto, T. Ogasawara, and A. Zelinsky, "Behavior recognition based on head pose and gaze direction measurement," in *IEEE IROS*, vol. 3, 2000.
- [10] T. Ohno and N. Mukawa, "A free-head, simple calibration, gaze tracking system that enables gaze-based interaction," in *Proceedings of the ETRA Symposium*. ACM, 2004.
- [11] F. Lu, Y. Sugano, T. Okabe, and Y. Sato, "Adaptive linear regression for appearance-based gaze estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, Oct 2014.
- [12] F. Lu, T. Okabe, Y. Sugano, and Y. Sato, "Learning gaze biases with head motion for head pose-free gaze estimation," *Image and Vision Computing*, vol. 32, 2014.
- [13] E. Murphy-Chutorian and M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, 2009.
- [14] X. Zabulis, T. Sarmis, and A. A. Argyros, "3D head pose estimation from multiple distant views," in *BMVC*, 2009.
- [15] I. Chamveha, Y. Sugano, D. Sugimura, T. Siriteerakul, T. Okabe, Y. Sato, and A. Sugimoto, "Head direction estimation from low resolution images with scene adaptation," *Computer Vision and Image Understanding*, vol. 117, 2013.
- [16] A. K. Rajagopal, R. Subramanian, E. Ricci, R. L. Vieri, O. Lanz, and N. Sebe, "Exploring transfer learning approaches for head pose classification from multi-view surveillance images," *International Journal of Computer Vision*, vol. 109, 2014.
- [17] Y. Yan, E. Ricci, R. Subramanian, G. Liu, O. Lanz, and N. Sebe, "A multi-task learning framework for head pose estimation under target motion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, 2016.
- [18] Z. Qin and C. R. Shelton, "Social grouping for multi-target tracking and head pose estimation in video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, 2016.
- [19] J. S. Stahl, "Amplitude of human head movements associated with horizontal saccades," *Experimental Brain Research*, vol. 126, 1999.
- [20] H. H. Goossens and A. Van Opstal, "Human eye-head coordination in two dimensions under different sensorimotor conditions," *Experimental Brain Research*, vol. 114, 1997.
- [21] R. Stiefelhagen and J. Zhu, "Head orientation and gaze direction in meetings," in *Human Factors in Computing Systems*, 2002.
- [22] S. Asteriadis, K. Karpouzis, and S. Kollias, "Visual focus of attention in non-calibrated environments using gaze estimation," *International Journal of Computer Vision*, vol. 107, 2014.
- [23] S. Ba and J.-M. Odobez, "Recognizing visual focus of attention from head pose in natural meetings," *IEEE Transactions on System Men and Cybernetics. Part B*, 2009.
- [24] S. Sheikh and J.-M. Odobez, "Recognizing the visual focus of attention for human robot interaction," in *Human Behavior Understanding Workshop*, 2012.
- [25] Z. Yucel, A. A. Salah, C. Mericli, T. Mericli, R. Valenti, and T. Gevers, "Joint attention by gaze interpolation and saliency," *IEEE Transactions on System Men and Cybernetics. Part B*, 2013.
- [26] M. J. Marin-Jimenez, A. Zisserman, M. Eichner, and V. Ferrari, "Detecting people looking at each other in videos," *International Journal of Computer Vision*, vol. 106, 2014.
- [27] K. Otsuka, J. Yamato, and Y. Takemae, "Conversation scene analysis with dynamic bayesian network based on visual head tracking," in *IEEE ICME*, 2006.
- [28] S. Duffner and C. Garcia, "Visual focus of attention estimation with unsupervised incremental learning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2015.
- [29] S. Sheikh and J.-M. Odobez, "Combining dynamic head pose-gaze mapping with the robot conversational state for attention recognition in human-robot interactions," *Pattern Recognition Letters*, vol. 66, 2015.
- [30] B. Massé, S. Ba, and R. Horaud, "Simultaneous estimation of gaze direction and visual focus of attention for multi-person-to-robot interaction," in *IEEE ICME*, Seattle, WA, Jul. 2016.
- [31] K. P. Murphy, "Switching Kalman filters," UC Berkeley, Tech. Rep., 1998.
- [32] D. Simon, "Kalman filtering with state constraints: a survey of linear and nonlinear algorithms," *Control Theory Applications, IET*, 2010.
- [33] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006.
- [34] V. Drouard, R. Horaud, A. Deleforge, S. Ba, and G. Evangelidis, "Robust head-pose estimation based on partially-latent mixture of linear regressions," *IEEE Transactions on Image Processing*, vol. 26, Jan. 2017.
- [35] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE CVPR*, vol. 1, 2001.
- [36] S.-H. Bae and K.-J. Yoon, "Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning," in *IEEE CVPR*, 2014.